An Opinion on Data Mining

PhD Student: Claus Gwiggner

3rd November 2005

Abstract

A purely data driven, assumption free approach for the discovery of patterns in data has a weak scientific base because of the underlying induction problem. Statistical inference builds its arguments on probabilistic models and not on the observed data. A class of probabilistic models are empirical models. They are not based on a theory of the logic of the system under study. Data mining can be seen as beeing concerned with the construction of empirical models. Then, inferences for predictions are possible but interpretation and insight into the phenomenon are highly critizized. Conclusion: without taking position to the problem of induction, interpretation of results will be critical.

Uncertainty, Inference, Data Mining and Models In this section we discuss the role of data mining in the construction of meaningful statements about observed phenomena. We look at different formulations of 'data mining' until the principles become clear. A purely descriptive analysis poses no particular question because any idea on how to display or summarize data may be useful for gaining intuition about a problem.

Data mining is sometimes described as the automatic extraction of 'knowledge' of a collection of data (eg. [KDD99]). Ideally, the methods should be free of assumptions on the data structure, so that models are the result of the algorithms and not assumptions of the scientist. Arguments against this idea can be found: without any assumptions of how the data might have been generated, '(...) no rational basis exists to generalize beyond the observed' [Mit97]. The philosophical dimension of the underlying induction problem will not be elaborated here (see e.g [Cha03]).

A more precise formulation of data mining is to consider the data as realizations of random variables and to 'infer' the properties of their probability distributions (e.g. [HTF03]). Inference because it involves making statements about observed phenomena. The theory of statistical inference (e.g. [Cha03], [Lin03]) adresses this problem. Different branches of this theory exist but one *central* assumption is that the mechanisms producing the observed data can be described by probabilistic models (e.g. [Cha03]), [Lin03]). The statements about the phenomenon are based on the assumption that the model is a good description of reality and not on the observations.

A simple example is the linear regression model which assumes that $Y_i \sim \mathcal{N}(\alpha_0 + \alpha_1 x_i, \sigma_\epsilon^2)$, i.e. Y is normally distributed with mean a linear function of x and constant variance σ_ϵ^2. Based on this assumption, uncertainty in the data can be quantified and, for example, precision of the estimated parameters can be inferred. An even simpler example is the following: Considering the data as independent realizations of random variables. Data mining often makes use of this minimal model in order to infer accuracy of prediction errors estimated by cross-validation [MB03]. Other 'inductive principles' than those in classical literature on statistical inference also exist [Sch04].

In this context, models can be broadly classified as *substantive* and *empirical*, depending on how much of the logic of the phenomenon is known (e.g.[CW98]). In this light, data mining can be further precised: as beeing concerned with building empirical models [HMS01]. They are successfully used in prediction tasks. When interpretation and understanding of the data is sought, their use is heavily critized [Bre01], [CW98].

References

[Bre01] L. Breiman. Two cultures in statistical modelling. With discussion. *Statistical Science*, 16-3, 2001.

[Cha03] S. K. Chatterjee. *Statistical Thought. A Perspective and History.* Oxford University Press, 2003.

[CW98] D.R. Cox and N. Wermuth. *Multivariate Dependencies. Models, Analysis and Interpretation.* Chapman and Hall/ CRC, 1998.

[HMS01] D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining.* MIT Press, 2001.

[HTF03] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction.* Springer Series in Statistics, 2003.

[KDD99] KDD. *Knowledge Discovery in Databases.* 1999.

[Lin03] J.K. Lindsey. *Parametric Statistical Inference.* Oxford University Press, 2003.

[MB03] J. Maindonald and J. Braun. *Data Analysis and Graphics Using R - An Example-Based Approach.* Cambridge University Press, 2003.

[Mit97] T. Mitchell. *Machine Learning.* Mc Graw Hill, 1997.

[Sch04] B. Schoelkopf. Statistische Lerntheorie und Empirische Inferenz. *Jahrbuch der Max-Planck-Gesellschaft*, 2004.